CLIMACAST: An Intelligent Weather Prediction System Based On Decision Tree Algorithms

[1]Abdulsalami, B.A., [2]Abdullateef, H.O. & [3]Adeola, L.O.
[1, 2, 3] Department of Mathematical & Computer Sciences
Fountain University
Osogbo, Osun State, Nigeria
E-mail: [1]basiratabdusalam@gmail.com, [2]abdullateefhamzah@ymail.com, [3]adeolalateefat19@yahoo.com

ABSTRACT

Weather prediction is often very difficult as a result of the unstable nature of the atmosphere. Until now, different approaches have been proposed in literatures by researchers, each with varying prediction accuracy. In this paper, we employed the use of data mining techniques for predicting weather. This was achieved by training with Decision Tree classification algorithms and Expectation-Maximization (EM) distribution-based clustering algorithms the meteorological data collected in specific time and location in Nigeria. The weather data that was used for the system includes temperature, pressure, humidity, wind and rainfall. In addition, a desktop application and web-based application, whose major logic was based on Decision Tree Algorithm, was developed for broadcast purposes. To make the solution feasible for more interactive usage irrespective of time and location, an android mobile interface was created. The system was implemented and deployed using Java programming technologies and WEKA 3.8 Machine Learning tools and libraries

Keywords: Decision Tree, C4.5 algorithm, Weather prediction, forecasting, clustering, data mining.

iSTEAMS Multidisciplinary Conference Proceedings Reference Format
Abdulsalami B.A., Abdullateef H.O. & Adeola L.O. (2019): CLIMACAST: An Intelligent Weather Prediction System Based On Decision Tree Algorithms. Proceedings of the 22nd iSTEAMS Multidisciplinary SPRING Conference. Aurora Conference centre, Osogbo, Nigeria. 17th – 19th December, 2019. Pp 101-118. www.isteams.net/spring2019. DOI - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V22N1P9

1. INTRODUCTION

Weather forecasting is the application of scientific and technological methods to predict the climatic condition or state of the atmosphere for a future time and location given location (Nazim & Ajith, 2013). The prediction of weather condition is essential for various applications such as climate monitoring, drought detection, pollution dispersal, etc. These are applicable and significant to sectors such as agriculture, production, energy industry, aviation industry, communication etc. For instance, in military operations, there is a considerable historical record of instances when weather conditions have altered the course of battles. (Prashant et al, 2017). In addition, users plan their daily routine with respect to possible weather conditions that may affect those activities. Since outdoor activities are severely curtailed by heavy rain and wind. The accurate prediction of weather conditions is found to be a herculean task due to the dynamic nature of atmosphere. Weather forecasting now relies on computer-based models that take many atmospheric factors such as: humidity, atmospheric pressure, wind, rainfall, outlook, into account.

Weather prediction had been known to be a complex and challenging task due to the unstable and chaotic nature of the atmosphere, the massive computational power required to solve the equations that describe the atmosphere, the error involved in measuring the initial conditions, and an incomplete understanding of atmospheric processes. Making an accurate prediction about meteorological characteristics has led to the use of several methods which vary in accuracy.

According to Siddharth and Roopa (Siddharth & Roopa, 2016), there are two methods to predict weather; the Empirical approach and Dynamical approach. The former depends on investigation of past chronicled information of forecast, which is gathered in meteorologist's environment and its relationship to an assortment of environmental variables over various parts of areas (Bhartkande & Huballi, 2016), and it is useful in predicting local scale weather if recorded cases are plentiful (Nazim O.B. *et al*, 2013), while in the later approach, expectations are produced by physical models; taking into account arrangement of conditions that anticipate the future climate figure. To foresee the climate by numeric means, meteorologist created air models that inexact the adjustment in temperature weight (Bhartkande & Huballi, 2016). The dynamical approach is useful to predict large scale weather phenomena and may not predict short term weather efficiently. Most weather prediction systems use a combination of both techniques.

Until now, different approaches have been proposed in literatures by researchers, each with varying prediction accuracy. In this paper, we proposed data mining techniques for weather prediction through the implementation of numerical methods using Decision Tree classification algorithms and Expectation-Maximization (EM) distribution-based clustering algorithms. Meteorological dataset collected in specific time and region such as; temperature, humidity, pressure, rainfall and wind speed were used to train and verify the effectiveness of these models. The system predicts weather based on parameters such as temperature, humidity, pressure and wind, using a software application developed in Java with effective graphical user interface. The administrator supplies the required data which the system then uses to predict weather based on correlation with previous data in the database and the classified training dataset model generated with the Waikato Environment for Knowledge Analysis (WEKA) Java library. A web-based application was developed for broadcast purposes. In order to make the solution feasible for more interactive usage at anywhere and at any time, an android mobile interface was created. The goal of the system is to achieve an intelligent, efficient, user friendly, interactive, automated, accurate, time and cost effective system.

The paper is organized as follows: Section 2 presents the review of relevant literatures related to this work, Section 3 shows the methodology adopted. Section 4 discusses the result of implementation of the system and then conclusions and future work are presented in Section 5.

2. LITERATURE REVIEW

Some very useful researches have been conducted in the area of weather forecasting, various mathematical, statistical and data mining approaches have been used by different researchers for building and validating prediction models on climatic condition. In 2009, E. G. Petre proposed the application of Classification and Regression Tree (CART) decision tree algorithm for weather prediction. The data collected is registered over Hong Kong. The data was recorded between 2002 and 2005. The data used for creating the dataset includes parameters such as; year, month, average pressure, relative humidity, clouds quantity, precipitation and average temperature. WEKA, open source data mining software, was used for the implementation of CART decision tree algorithm. The decision tree, results and statistical information about the data are used to generate the decision model for prediction of weather.

A decision tree was produced as an output and its performance was calculated using evaluation metrics which included parameters like precision, accuracy, FP rate, TP rate, F-measure, and ROC Area.

Olaiya and Adeyemo, in 2012, investigated the use of data mining techniques in predicting maximum temperature, rainfall, evaporation and wind speed. C4.5 decision tree algorithm and artificial neural networks are used for prediction. The meteorological data is collected between 2000 and 2009 from the city Ibadan, Nigeria. A data model for the meteorological data is developed and is used to train the classifier algorithms. The performance of each algorithm is compared with the standard performance metrics and the algorithm with the best result is used to generate classification rules for the mean weather variables. A predictive neural network model is also developed for weather prediction and the results are compared with the actual weather data for the predicted period. The results shows that given enough training data, data mining technique can be efficiently used for weather prediction and climate change studies.

In 2013, Gaurav and Sunil, proposed an artificial neural network method for the prediction of weather for future in a given location. Back Propagation Neural Network is used for initial modeling. Then Hopfield Networks are fed with the result outputted by BPN model. The attributes include temperature, humidity and wind speed. Three years data of weather is collected comprising of 15000 instances. The prediction error is very less and learning process is quick. This can be considered as an alternative to the traditional meteorological approaches. Both algorithms are combined effectively. It is able to determine non- linear relationship that exists between the historical data attributes and predicts the weather in future.

P. S. Mohod *et.al*, 2013 stated that rainfall and weather forecasting is most challenging problem in around the world in the agricultural field. Their work describes the algorithms related to data mining which are used for the determination. Neural network (NN), random forest, CART, support vector machine (SVM) and K-nearest neighbor. They applied frequent mining on the available dataset for discovering frequent pattern. Frequent item will find from dataset on parameters like temperature, humidity, wind. These algorithms applied on the last five year rainfall dataset.

In 2013, Chandar S. suggested the application of the algorithm ID3 developed by Ross Quinlan, which is a simple decision tree supervised learning algorithm. To test each attribute at every tree node, the decision tree was constructed by employing a top - down, greedy search algorithm. This work introduced a new metric named information gain to select the attribute which is useful for classifying the given set. Divya and Jawahar, in 2014, described the capabilities of various algorithms in predicting several weather phenomena such as temperature, windy, humidity, rainfall. They concluded that major techniques like decision trees, artificial neural networks, clustering and regression algorithms are suitable to predict weather phenomena. Which shows that decision trees and k-means clustering are best suited data mining techniques for this application.

In 2013, Valmik B. N. and B.B. Meshram (2013) proposed a model for predicting the weather data based on classification technique and considered several attributes such as wind pressure, humidity, vapor, wind speed and cynical results obtained. The results showcased good accuracy by correlating the above parameters.
Z. U. Khan and M. Hayat, in 2014, implemented various data mining techniques for prediction of weather forecasting including different classifications like K-Nearest Neighbour, Decision Trees and Naïve Bayes. Decision Trees has achieved quite promising performance among the algorithms. Among the classification algorithms, decision tree achieved promising results compared to other algorithms. The predicted outcomes are 82.62% of accuracy.

In 2015, Ashwini M. and, Jadhawar B.A. applied Artificial Neural Networks and Decision Tree algorithms in prediction of weather. ANN finds a relationship between the weather attributes and builds a complex model, whereas C5 decision tree learns the trend of data and accordingly builds a classifier tree that can be used for prediction. Siddharth S.B., and Roopa G. H., in 2016, implemented Decision tree algorithm for classifying weather parameters such as maximum temperature, minimum temperature in terms of the day, month and year. The data used from wounder ground weather site between 2012 and 2015 from different cities. The results showed how these parameters have influenced the weather observed in these months over the study period.
In 2016, Liu et al. developed Deep Convolutional Neural Network (CNN) classification system. The work demonstrated the usefulness of Deep Learning technique for tackling climate pattern detection problems. Coupled with Bayesian based hyper-parameter optimization scheme, their deep CNN system achieves 89\%-99\% of accuracy in detecting extreme events.

S. Karthick and D. Malathi, in 2018 did a comparative analysis between C4.5 Decision Tree (J48) and Random Forest algorithm with dataset comprising of nine weather parameters collected over a period of two years. Though the result of both the algorithms was found to be relatively good as they fall in the category of recommended algorithms for classification and weather prediction problems, yet Random Forest proved to be better than the C4.5 Decision Tree. Where C4.5 achieved an accuracy of 82.4%, Random Forest was able to secure 87.1%accuracy proving it to be better. The accuracy was obtained using 10 fold cross validation keeping the over fitting problem in mind.

3. METHODOLOGY

For the purpose of this work, meteorological dataset collected from the meteorological lab of Fountain University Osogbo such as: temperature, humidity, pressure, rainfall and wind speed were used to train and verify the effectiveness of the algorithms. The two different supervised learning algorithms from Waikato Environment for Knowledge Analysis (WEKA) software were used in building the predictive models. The learning techniques used are: Decision Tree classification algorithms and Expectation-Maximization (EM) distribution-based clustering algorithms. To make the system feasible for more interactive usage, weather prediction software implementing the decision tree algorithm was constructed as well as a web-based application, using NetBeans; and android mobile application was implemented using Android Studio. *The flow chart of the system is depicted in Fig. 1.*

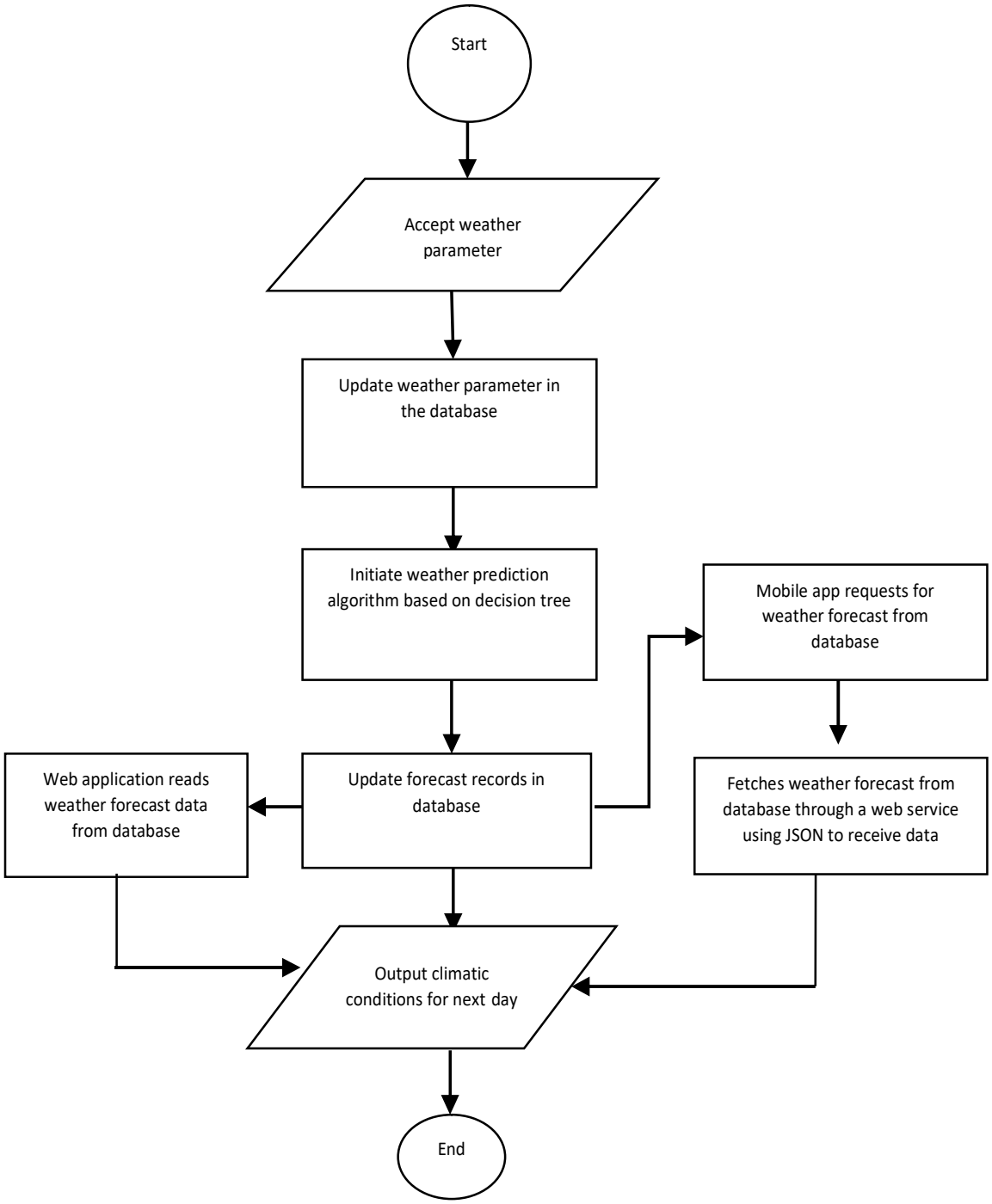

Fig. 1: System flowchart

3.1. Data Collection

The meteorological data used for this system was obtained from the meteorological lab of Fountain University, Osogbo, Osun State. The data covered three (3) years period from 2014 to 2016. The following procedures were applied to the collected data: Data cleaning, Data selection, Data transformation, and Data mining. Data Cleaning: In this stage, a consistent format for the data model is developed, which takes care of missing data, finds duplicated data, and weeds out bad data. Data Selection: Data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset has ten (6) attributes, their type and description are presented in Table 1, and the analysis of the numeric values presented in Table 2. The *Analysis of the numeric values presented in Table 2.*

Table 1: Attributes of Meteorological Dataset

Attribute	Type	Description
Month	Numerical	Month considered
Wind Speed	Numerical	Wind run in km
Temperature	Numerical	The average temperature
Rainfall	Nominal	Whether it rains or not
Humidity	Numerical	Humidity
Pressure	Numerical	Pressure

Table 2: Analysis of Weather Data

No	Variable	Min	Max	Mean	SD	Missing Values
1	Wind speed	79.33	188.78	134.913	23.696	0%
2	Humidity	1.7	10.9	4.128	1.898	0%
3	Pressure	1.5	7.9	5.07	1.756	0%
4	Temp	21.1	30.9	23.157	1.35	0%
5	Rainfall	No	yes	-	-	0%
6	Month	1(Jan)	12(Dec)	-	-	-

Data Transformation: The selected data is transformed into forms appropriate for data mining. The datasets are normalized and the data file was saved in attribute relation file format (arff) file format for use with Waikato Environment for Knowledge Analysis (WEKA); a data mining software. Data mining: This stage involved several phases. At each phase, the algorithms are used to analyze the dataset. The C4.5 Decision Tree classifier algorithm which is implemented as J48 in WEKA was used to analyze the dataset.

3.2. Description of the Data Mining Tools.

The data mining tools used for building the models are described below:

(i) C4.5 Decision Tree algorithm / J48: Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes. They are not only commonly used specifi-cally in decision analysis as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. In prediction analysis, decision tree models are commonly used to examine the data and to induce the tree and its rules that will be used to make predictions. Predictive models maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). There are many varieties of decision-tree al-gorithms. Among them is C4.5. It is a decision tree algorithm developed by Ross Quinlan. It is an ex-tension of Iterative Dichotomizer 3(ID3) algorithm. It can be used for classification, and often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. J48 is an open source java implementation of the C4.5 algorithm in the WEKA data mining tool.

(ii) EM Cluster Analysis: Expectation-Maximization (EM), a type of distribution-based clustering algo-rithm was used to analyse and group each parameter in our meteorological dataset in such a way that objects in the same group are similar so as to study and establish the patterns of progression or re-gression in our dataset for predicting future values of these parameters.

3.3. The Simulation Environment

The simulation environment used for the prediction model is Waikato Environment for Knowledge Anal-ysis (WEKA). WEKA is a popular suite of machine learning software written in Java, and developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, to-gether with a graphical user interface (GUI) for easy access to these functions. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

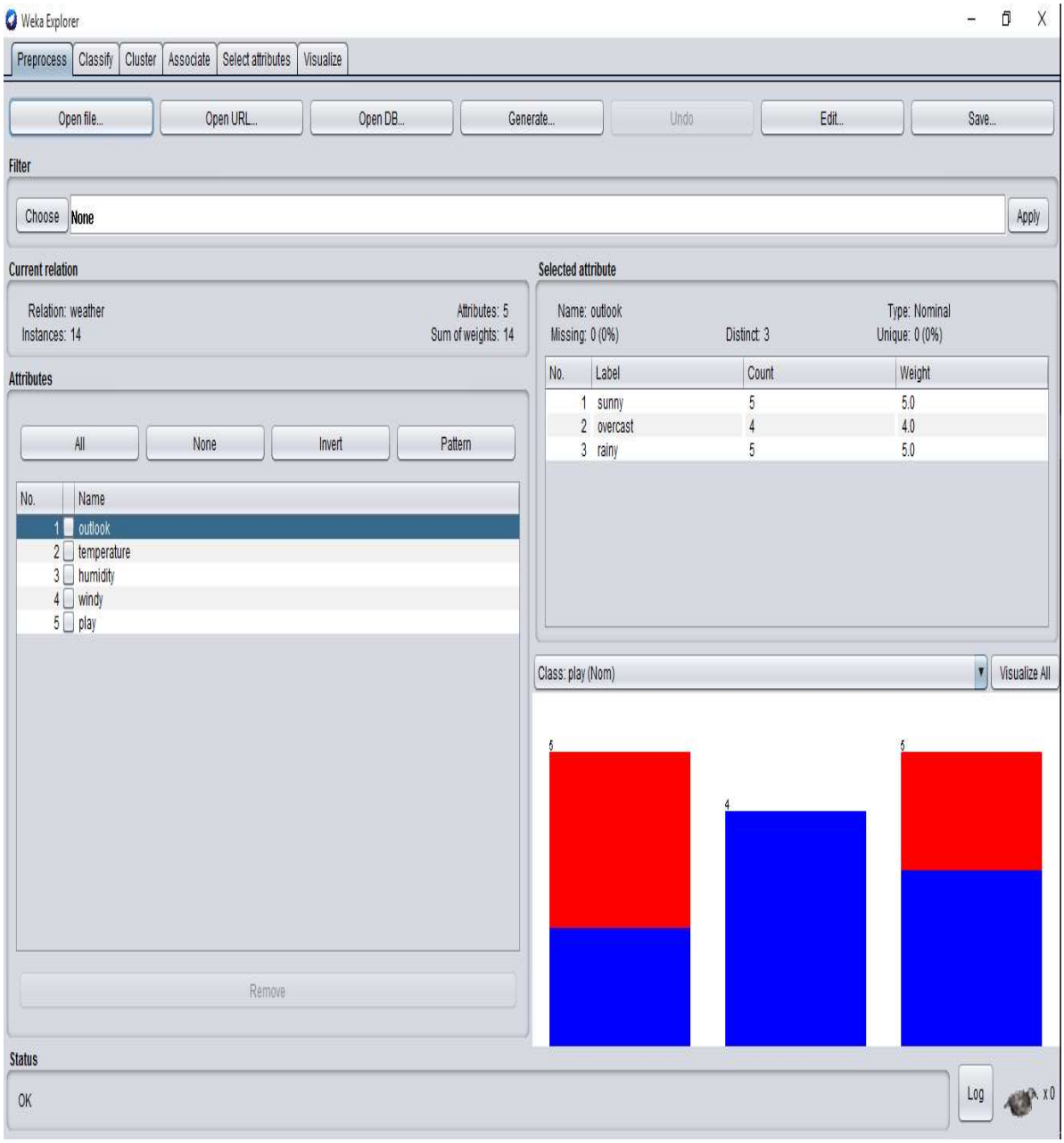

Fig. 5: Pre-processing imported data in the WEKA Simulation environment.

In our proposed system, WEKA was used to preprocess our meteorological data and its J48 classifier was used to classify our decision tree. The required data are supplied to the system and pre-processed, as shown in Figure 5.